VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Konstrukce intervalů spolehlivosti bootstrapovou metodou

The construction of bootstrap confidence intervals

2016

Jan Sabel

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

Zadání bakalářské práce

Student: **Jan Sabel**

Studijní program: B2647 Informační a komunikační technologie

Studijní obor: 1103R031 Výpočetní matematika

Téma: Konstrukce intervalů spolehlivosti bootstrapovou metodou
The construction of bootstrap confidence intervals

Jazyk vypracování: čeština

Zásady pro vypracování:

Metoda bootstrap patři mezi tzv. intenzivni počitačove metody pro statistickou analýzu dat. Cílem práce je popis metody bootstrap, jejích vlastností a užití pro konstrukci intervalů spolehlivosti. Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Postup práce:
1. Metoda bootstrap.
2. Užití bootstrap metody pro konstrukci intervalů spolehlivosti.
3. Srovnání klasických intervalových odhadů a bootstrapových intervalových odhadů (srovnání pokrytí a délky intervalových odhadů na základě simulací).

Seznam doporučené odborné literatury:

1. PRÁŠKOVÁ Z. (2004), Metoda bootstrap, Sborník ROBUST 2004.
2. DICICCIO T. J., EFRON B. (1996), Bootstrap Confidence Intervals, Statistical Science, Volume 11, No 3, pg. 189 – 228.

Formální náležitosti a rozsah bakalářské práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí bakalářské práce: **Ing. Martina Litschmannová, Ph.D.**

Datum zadání: 01.09.2015
Datum odevzdání: 15.07.2016

doc. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

VYSOKÁ ŠKOLA BÁŇSKÁ-TECHNICKÁ UNIVERZITA Fakulta elektrotechniky a informatiky 1D16 OSTRAVA

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

„Prohlašuji, že jsem tuto bakalářskou/diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.“

V Ostravě dne 15. 7. 2016

Jan Sabel

Mnohokrát děkuji vedoucí práce, paní Ing. Martině Litschmannové, Ph.D., za neskutečnou ochotu a trpělivost projevenou při vzniku této práce. Dále děkuji rodičům, kteří mi byli nedocenitelnou oporou při psaní této práce.

Abstrakt

Cílem práce je seznámit čtenáře s metodou bootstrap a jejím použitím pro konstrukci intervalů spolehlivosti. Nejprve je vysvětlen princip metody bootstrap, následně její použití při tvorbě intervalu spolehlivosti. V poslední kapitole je metoda porovnávána s klasickým způsobem určení intervalu spolehlivosti.

Klíčová slova: Bootstrap, bootstrapový výběr, interval spolehlivosti, konstrukce intervalu spolehlivosti

Abstract

The goal of thesis is to introduce a bootstrap method and its usage for construction of confidence intervals to reader. First of all is explained a principle of the bootstrap method. Afterwards is explained use of the bootstrap method in constructing of confidence intervals. In last chapter is bootstrap compared to a classic way of constructing confidence intervals.

Keywords: Bootstrap, bootstrap sample, confidence interval, construction of confidence interval

Seznam použitých zkratek a symbolů

Ω	–	Prostor elementárních jevů
A	–	Náhodný jev
ω	–	Elementární jev
$\sum$	–	Jevové pole
$P(A)$	–	Pravděpodobnost jevu A
X	–	Náhodná veličina
$F(x)$	–	Distribuční funkce náhodné veličiny X
$f(x)$	–	hustota pravděpodobnosti distribuční funkce $F(x)$
$E(X)$	–	Střední hodnota náhodné veličiny X
$D(X)$	–	Rozptyl náhodné veličiny X
Θ	–	Reálný parametr daného rozdělení pravděpodobnosti
$\bar{X}$	–	Výběrový průměr
S^2	–	Výběrový rozptyl
S	–	Výběrová směrodatná odchylka
μ	–	Střední hodnota
σ^2	–	Rozptyl
σ	–	Směrodatná odchylka
$N(\mu, \sigma^2)$	–	Normální rozdělení s parametry μ a σ^2
$N(0, 1)$	–	Normované normální rozdělení
$LN(\mu, \sigma)$	–	Logaritmicko-normální rozdělení s parametry μ a σ
$Exp(\lambda)$	–	Exponenciální rozdělení s parametrem λ
$1 - \alpha$	–	Spolehlivost odhadu
I	–	Pomocný identifikátor používaný pro testování intervalových odhadů
C	–	Pravděpodobnost pokrytí
n	–	Rozsah výběru
IO	–	intervalový odhad

Obsah

Seznam obrázků

1 Úvod

S myšlenkou metody bootstrap přišel v roce 1979 Bradly Efron. V té době se jednalo o věc až revoluční, neboť namísto algebraických výpočtů využívala počítačovou simulaci. Velký rozvoj metody však nastal až s rozvojem počítačů a internetu. V dnešní době, kdy se nové poznatky mohou rozšířit po celém světě okamžitě a kdy je více-jádrový procesor běžnou součástí i domácích počítačů, nabírá bootstrap na stále větší důležitosti. Jde o velmi univerzální a jednoduchou metodu použitelnou v mnoha oborech statistiky. V této bakalářské práci se zaměříme na použití metody bootstrap pro konstrukci intervalových odhadů a prozkoumáme jejich efektivitu. Nejprve si připomeneme některé pojmy ze statistiky.

2 Základní pojmy

2.1 Jevy

Náhodný pokus je realizací počátečních podmínek, které jsou neměnné. Jeho výsledky jsou, při stejných podmínkách, různé. Množinu všech výsledků nazveme **základní prostor** (prostor elementárních jevů) a značíme jí Ω. Jednotlivé výsledky nazveme **náhodné jevy** a značíme je A. Náhodný jev je libovolná podmnožina základního prostoru Ω. **Jevem opačným** k jevu A rozumíme jev, který nastane právě tehdy když nenastane jev A. Značíme jej $\bar{A}$. Speciální případ náhodného jevu je **elementární jev**, který značíme ω.

Definice 2.1 *Elementární jev ω je takový náhodný jev, pro který neexistují jevy B a C různé od ω takové, že $\omega = B \cup C$.*

Definice 2.2 *Jevové pole $\sum$ na základním prostoru Ω je množina náhodných jevů, pro kterou platí:*

1. *Pro každý náhodný jev $A \in \sum$ je opačný jev $\bar{A} \in \sum$.*
2. *Pro každou posloupnost náhodných jevů $A_i \in \sum, i = 1, 2, ...$ je $\bigcap_{i=1}^{\infty} A_i \in \sum$.*

Definice 2.3 *Mějme náhodný jev A z jevového pole $\sum$. Pak funkci $P(A)$ nazveme pravděpodobnostní a platí pro ni:*

1. *Pro každý náhodný jev A z jevového pole $\sum$ je $0 \leq P(A) \leq 1$.*
2. *$P(\Omega) = 1$.*
3. *Pro každou posloupnost disjunktních jevů A_i z jevového pole $\sum$ platí:*
   $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

*Uspořádanou trojici $(\Omega, \sum, P)$ nazveme **pravděpodobnostní prostor**.*

2.2 Náhodné veličiny

Náhodná veličina X je funkce, která elementárním jevům ω přiřazuje reálné čísla x. Jedná se tedy o zobrazení $X : \Omega \to \mathbb{R}$. X je náhodná veličina na pravděpodobnostním prostoru $\sum$ právě tehdy, když pro každé reálné číslo $x1$ platí: $\{\omega : X(\omega) < x\}$. Množinu všech hodnot náhodné veličiny X nazýváme **základní soubor** (populace). Známe 2 druhy náhodných veličin - diskrétní a spojité. Rozdělení náhodné veličiny je dáno její distribuční funkcí.

Rozdělení pravděpodobnosti lze interpretovat jako zobrazení $\omega \to \mathbb{R}$, které elementárním jevům ω přiřazuje reálné číslo, které charakterizuje pravděpodobnost $X(\omega) \in M; M$ je libovolná podmnožina $\mathbb{R}$.

Definice 2.4 *Funkci $F(x)$ nazveme **distribuční funkcí** náhodné veličiny X, platí-li, že každému reálnému $x \in (-\infty, +\infty)$ přiřazuje pravděpodobnost toho, že $X < x, tj. F(x) = P(X < x)$.*

Poznámka 2.1 Distribuční funkce je neklesající, zleva spojitá a platí pro ni $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$. Může mít nejvýše spočetně bodů nespojitosti.

Je-li rozdělení diskrétní, můžeme označit body nespojitosti jako x_1, x_2.... Pak platí $p_k = P(X = x_k) = lim_{x\to x_k} F(x) - F(x_k)$..

Je-li rozdělení spojité, pak existuje funkce $f(x)$ taková, že platí: $F(x) = \int_{-\infty}^{x} f(t)dt$. Funkci $f(x)$ nazýváme **hustotou pravděpodobnosti** distribuční funkce $F(x)$. Funkce $F(x)$ tedy musí být absolutně spojitá. Jelikož je $F(x)$ neklesající, $f(x) \geq 0$ skoro všude.

Nebude-li uvedeno jinak, v celé bakalářské práci se bude pracovat pouze se spojitým rozdělením.

2.3 Číselné charakteristiky

Vlastnosti náhodné veličiny X jsou popsány jejími **číselnými charakteristikami**. Mezi nejdůležitější patří střední hodnota, rozptyl a směrodatná odchylka. Pro spojitou náhodnou veličinu X je určíme jako:

Definice 2.5 *Střední hodnotu $E(X)$ pro náhodnou veličinu X určíme jako:*
$E(X) = \int_{-\infty}^{\infty} xf(x)dx$.

Definice 2.6 *Rozptyl $D(X)$ náhodné veličiny X určíme jako:*
$D(X) = E([X - E(X)]^2 = E(X^2) - [E(X)]^2$.

Definice 2.7 *Směrodatnou odchylku $\sigma(x)$ náhodné veličiny X určíme jako:*
$\sigma(x) = \sqrt{D(X)}$.

Při hledání intervalového odhadu se setkáme s kvantily.

Definice 2.8 $100p\%$ ***kvantil*** x_p *je číslo, pro které platí, že pravděpodobnost, že náhodná veličina bude mít hodnoty menší než x_p je p. Platí tedy:*
$P(X < x_p) = F(X_p) = p$.

Poznámka 2.2 Je-li X spojitá náhodná veličina, pak $P(X < x_p) = P(X \leq x_p)$.

2.4 Náhodný výběr

Opakováním náhodného pokusu získáváme náhodné veličiny X_i s distribuční funkcí $F(x, \Theta)$, kde Θ značí reálný parametr daného rozdělení pravděpodobnosti. Jednotlivé veličiny můžeme uspořádat do **náhodného vektoru** $X = (X_1, X_2, ..., X_n)$. Jednotlivé složky náhodného vektoru X musí být nezávislé náhodné veličiny X_i, které mají stejné rozdělení pravděpodobnosti. Pak náhodný vektor $(X_1, ...X_n)$ nazveme **náhodným výběrem** z náhodné veličiny X. Číslo n nazveme **rozsah náhodného výběru**. Realizací $(X_1, ...X_n)$ získáme číselný vektor $x = (x_1, ..., x_n)$. Tento vektor nazýváme **pozorovaná hodnota náhodného výběru**.

Funkci náhodného výběru $T(X_1, ..., X_n)$ nazveme **výběrová charakteristika**. Hodnotu $t = T(x_1, ..., x_n)$ nazveme **empirická charakteristika**. Nejdůležitější výběrové charakteristiky jsou:

1. **Výběrový průměr**, který značíme $\bar{X}$ a určíme jako:
   $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
2. **Výběrový rozptyl**, který značíme S^2 a určíme jako:
   $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$.
3. **Výběrová směrodatná odchylka**, kterou značíme S a určíme jako
   $S = \sqrt{S^2}$.

Pro výběrové charakteristiky platí následující důležité vlastnosti:

Mějme náhodnou veličinu X se střední hodnotou $E(X)$, rozptylem $D(X)$ a rozsahem náhodného výběru n. Pak platí:

1. $E(\bar{X}) = E(X)$.
2. $D\bar{X} = \frac{D(X)}{n}$, $\sigma(\bar{X}) = \frac{\sigma(X)}{\sqrt{n}}$, $E(S^2) = D(X)$.

Poznámka 2.3 Je zřejmé, že $\lim_{n\to\infty} \frac{D(X)}{n} = 0$. To znamená, že pro $n \to \infty$ rozptyl průměru konverguje k nule, tj. $D(\bar{X}) \to 0$.

Hodnoty empirických charakteristik $t = T(x_1, ..., x_n)$ jsou realizacemi náhodné veličiny, pro různé náhodné výběry mají různé hodnoty.

2.5 Rozdělení pravděpodobnosti

Existuje mnoho různých rozdělení pravděpodobnosti, uvedeny budou pouze ty, se kterými se v této bakalářské práci bude dále pracovat.

2.5.1 Normální rozdělení

Normální (Gaussovo) rozdělení značíme $N(\mu, \sigma^2)$, kde μ značí střední hodnotu a σ^2 značí rozptyl . Hustotu pravděpodobnosti normálního rozdělení určíme následovně:

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{(x-\sigma)^2}{2\sigma^2}], x \in (-\infty, \infty)$.

Speciálním případem normálního rozdělení je $N(0; 1)$. Toto rozdělení nazýváme **normované normální rozdělení**.

2.5.2 Logaritmicko-normální rozdělení

Logaritmicko-normální rozdělení s parametry μ a σ značíme $LN(\mu, \sigma)$. Jedná se o rozdělení spojité náhodné veličiny X takové, že $ln(X)$ má normální rozdělení s parametry μ (střední hodnota) a σ (směrodatnou odchylka). Charakteristiky logaritmicko-normálního rozdělení jsou:

- střední hodnota $E(X) = e^{\mu+\sigma^2/2}$,
- rozptyl $D(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$.

Hustotu pravděpodobnosti logaritmicko-normálního rozdělení určíme následovně: $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(ln(x)-\sigma)^2}{2\sigma^2}}$.

2.5.3 Exponenciální rozdělení

Exponenciální rozdělení značíme $Exp(\lambda)$, kde λ značí **parametr exponenciálního rozdělení**, $\lambda > 0$. Charakteristiky exponenciálního rozdělení jsou:

- střední hodnota $E(X) = \frac{1}{\lambda}$,
- rozptyl $D(X) = \frac{1}{\lambda^2}$.

Hustota pravděpodobnosti exponenciálního rozdělení s parametrem $\lambda > 0$ má tvar: $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$

2.6 Odhady parametrů

Odhadem T parametru Θ rozumíme statistiku $T(X_1, X_2, ..., X_n)$.

2.6.1 Bodový odhad

Bodový odhad parametru Θ používáme především v situacích, kdy hodnotu parametru potřebujeme pro další výpočty. Bodovým odhadem parametru Θ nazveme hodnotu $t = T(x_1, x_2, ..., x_n)$ parametru T. Aby byl bodový odhad **věrohodný**, musí splňovat následující vlastnosti:

- nestrannost (nevychýlenost)
- vydatnost (eficience)
- konzistence

Poznámka 2.4 Řekneme, že T je **nestranný odhad (nevychýlený)**, jestliže platí $E(T) = \Theta$. Jestliže rovnost neplatí, nazýváme T **vychýlený odhad**.

Mějme m nestranných bodových odhadů parametru Θ $(T_1, T_2, ..., T_m)$. Pak **nejlepším nestranným eficientním odhadem parametru** Θ nazveme odhad s nejmenším rozptylem $D(T_i), 0 < i \leq m$.

Bodový odhad parametru Θ nazveme **konzistentní**, jestliže s rostoucím rozsahem výběru n se bodový odhad zpřesňuje. Tj. platí-li:

- $\lim_{n\to\infty} E(T_n) = \Theta$,
- $\lim_{n\to\infty} D(T_n) = 0$.

Poznámka 2.5 Přesnost bodového odhadu

Hodnota bodového odhadu je obyčejně mírně odlišná od skutečné hodnoty populace. Mějme nezkreslený bodový odhad T parametru Θ. Pak velikost odlišnosti $(T - \Theta)$ určuje **výběrová chyba**. Ta určuje velikost chyby na základě jednoho bodového odhadu. Měřítko chyby je směrodatná odchylka $\sigma_t = \sqrt{D(T)} = \sqrt{E(T-\Theta)^2}$, kterou někdy nazýváme **střední kvadratická chyba odhadu**.

2.6.2 Intervalový odhad

Při praktických aplikacích často stanovujeme odhad parametru pomocí intervalového odhadu. **Intervalový odhad parametru** Θ je interval $\langle t_d, t_h \rangle$, ve kterém parametr Θ leží s předem určenou spolehlivostí (pravděpodobností) $(1 - \alpha)$.

Definice 2.9 ***Interval spolehlivosti*** *(konfidenční interval) pro parametr* Θ *se spolehlivostí* $1 - \alpha, \alpha \in \langle 0; 1 \rangle$*, je taková dvojce statistik* (T_D, T_H)*, pro kterou platí:*

$P(T_D \leq \Theta \leq T_H) = 1 - \alpha$.

Spolehlivost odhadu $1 - \alpha$ udává, že při opakovaných výběrech s konstantním rozsahem m z dané populace bude přibližně $100 \cdot (1 - \alpha)\%$ intervalových odhadů (IO) obsahovat skutečnou hodnotu odhadovaného parametru Θ.

Číslo α nazýváme **hladina významnosti**. S rostoucí spolehlivostí odhadu $1 - \alpha$ klesá hladina významnosti α.

Poznámka 2.6 V praxi spolehlivost odhadu $1 - \alpha$ volíme nejčastěji $90\%, 95\%$ nebo 99%. Je patrné, že čím je spolehlivost odhadu bližší 1 (100%), tím širší interval je. Chceme-li interval zúžit, NIKDY bychom neměli snižovat spolehlivost odhadu, jelikož tím snižujeme informativní hodnotu odhadu. Snížení šířky je možno dosáhnout zvýšením rozsahu výběru n. S rostoucím rozsahem výběru se úměrně zmenšuje šířka intervalových odhadů $\sqrt{(n)}$-krát.

Poznámka 2.7 Nebude-li uvedeno jinak, všechny příklady i teoretické texty používají spolehlivost odhadu $1 - \alpha = 0.95$ (95%).

Intervalové odhady mohou být jednostranné nebo oboustranné.

Jednostranné intervaly spolehlivosti

Jednostranné intervaly spolehlivosti konstruujeme tehdy, je-li pro nás důležitá pouze jedna mez. U jednostranných intervalů spolehlivosti je udávána pouze dolní mez (T_D) nebo pouze horní mez (T_H).

- U **levostranného intervalu spolehlivosti** je udávána pouze dolní mez T_D a platí pro něj:

  $P(\Theta \geq T_D) = 1 - \alpha$.

- U **pravostranného intervalu spolehlivosti** je udávána pouze horní mez T_H a platí pro něj:

  $P(\Theta \leq T_H) = 1 - \alpha$.

Oboustranné intervaly spolehlivosti

Potřebujeme-li znát obě meze odhadu (dolní T_D i horní T_H), zkonstruujeme **oboustranný interval spolehlivosti**:

$P(T_D \leq \Omega \leq T_H) = 1 - \alpha$.

Poznámka 2.8 Obvykle se meze intervalu spolehlivosti určují tak, aby platilo, že pravděpodobnost, že parametr populace Θ leží pod dolní mezí T_D byla stejná jako pravděpodobnost, že parametr populace Θ leží nad horní mezí T_D a byla rovna $\frac{\alpha}{2}$. Platí tedy:

$P(\Theta < T_D) = P(\Theta > T_H) = \frac{\alpha}{2}$.

Sestrojení intervalového odhadu

Všechny způsoby konstrukce intervalů spolehlivosti uvedené v této kapitole předpokládají, že data pocházejí z normálního rozdělení, nebo pro ně platí centrální limitní věta.

Věta 2.1 ***Centrální limitní věta*** *Nechť:*

- $X_1, X_2, ...X_n$ *jsou nezávislé, stejně rozdělené náhodné veličiny,*
- $E(X_i) = \mu_x$,
- $D(X_i) = \sigma_x^2; D(X_i) < \infty$.

Pak výběrový průměr $\hat{x}$ má při dostatečně velkém počtu pozorování přibližně normální rozdělení, ať už X_i pocházejí z libovolného rozdělení. Tzn. platí:

$\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$.

Věta 2.2 ***Důsledek centrální limitní věty:*** *Označme součet náhodných veličin S_n : $S_n = \sum_{i=1}^{n} X_i$.*

Pak pro dostatečně velká n (za dostatečně velké n považujeme $n > 30$) platí:

$S_n \sim N(n\mu, n\sigma^2)$.

Intervalový odhad střední hodnoty normálního rozdělení

Lze dokázat, že nejlepší **bodový odhad** střední hodnoty μ je $\bar{x}$. **Intervalový odhad** střední hodnoty μ se určuje jinak známe-li rozptyl σ^2 resp. směrodatnou odchylku σ populace základního souboru a jinak, pokud rozptyl σ^2 resp. směrodatnou odchylku σ neznáme.

Mějme náhodnou veličinu X, která má normální rozdělení. Pak lze dokázat, že intervalové odhady pro známou, resp. neznámou směrodatnou odchylku σ určíme následovně.

Intervalový odhad střední hodnoty μ pro známou směrodatnou odchylku σ

Pro náhodnou veličinu X známe její rozptyl σ^2 a hledáme intervalový odhad její střední hodnoty. Náhodným výběrem určíme vzorek z populace o rozsahu n a průměru $\bar{x}$. Označme $1 - \frac{\alpha}{2}$ kvantil normované normální náhodné veličiny $z_{1-\frac{\alpha}{2}}$. Intervalové odhady střední hodnoty μ se spolehlivostí $1-\alpha$ při známém rozptylu σ^2 určíme následovně:

- **Oboustranný intervalový odhad střední hodnoty** μ se spolehlivostí $1-\alpha$ se známým rozptylem σ^2 se určí jako:

  $\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}; \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \rangle$.

- **Levostranný intervalový odhad střední hodnoty** μ se spolehlivostí $1-\alpha$ se známým rozptylem σ^2 je dán dolní mezí (střední hodnota μ je větší než dolní mez se spolehlivostí $1-\alpha$). Ta se určí jako:

  $\langle \bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \rangle$.

- **Pravostranný intervalový odhad střední hodnoty** μ se spolehlivostí $1-\alpha$ se známým rozptylem σ^2 je dán horní mezí (střední hodnota μ je menší než dolní mez se spolehlivostí $1-\alpha$). Ta se určí jako:

  $\langle \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \rangle$.

Poznámka 2.9 V předchozích definicích jsou použity $z_p 100p\%$ kvantily normovaného normálního rozdělení. Tyto kvantily lze dohledat ve statistických tabulkách nebo určit pomocí statistických počítačových programů.

Určit intervalový odhad pomocí výše uvedeného postupu lze i v případě, že směrodatnou odchylku neznáme, ale máme dostatečně velký výběr ($n \geq 30$). V tomto případě nahradíme směrodatnou odchylku σ výběrovou směrodatnou odchylkou s.

Intervalový odhad střední hodnoty μ pro neznámou směrodatnou odchylku σ

Náhodným výběrem určíme vzorek z populace o rozsahu n, průměru $\bar{x}$ a výběrovou směrodatnou odchylkou s. Označme $1-\frac{\alpha}{2}$ kvantil Studentova rozdělení $t_{1-\frac{\alpha}{2}}$. Intervalové odhady střední hodnoty μ se spolehlivostí $1-\alpha$ při neznámém rozptylu σ^2 určíme následovně:

- **Oboustranný intervalový odhad střední hodnoty** μ se spolehlivostí $1-\alpha$ s neznámým rozptylem σ^2 se určí jako:

  $\langle \bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}; \bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}} \rangle$.

- **Levostranný intervalový odhad střední hodnoty** μ se spolehlivostí $1-\alpha$ s neznámým rozptylem σ^2 je dán dolní mezí (střední hodnota μ je větší než dolní mez se spolehlivostí $1-\alpha$). Ta se určí jako:

  $\langle \bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha} \rangle$.

- **Pravostranný intervalový odhad střední hodnoty** μ se spolehlivostí $1-\alpha$ s neznámým rozptylem σ^2 je dán horní mezí (střední hodnota μ je menší než horní mez se spolehlivostí $1-\alpha$). Ta se určí jako:

  $\langle \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha} \rangle$.

3 Bootstrap

Metodu bootstrap představil B. Efron v roce 1979. Ve světě statistiky se rychle rozšířila a usadila. Dnes se jedná o velmi důležitou metodu, která má použití v různých oblastech statistiky. V této práci se zaměříme na použití pro konstrukci intervalů spolehlivosti. Nejprve si však vysvětlíme základní principy metody.

Při studii náhodného výběru je často statistikovým úkolem zjistit jeho charakteristiky a pomocí nich vyvodit parametry náhodné veličiny, ze které studovaný náhodný výběr pochází. Charakteristiky náhodného výběru někdy nazýváme výběrové statistiky. Mezi základní statistiky, které náš zajímají, patří **výběrový průměr**,**výběrový medián** a **výběrová směrodatná odchylka**. Je zřejmé, že pro různé náhodné výběry se budou charakteristiky (např. výběrový průměr) lišit. Zajímá nás, jak velká je celková variabilita výběrových statistik, abychom mohli určit přesnost. Rozdělení pravděpodobnosti všech možných hodnot statistiky, které mohou být zjištěny z náhodných výběrů, nazveme **výběrové rozdělení**. Velkou výhodou metody bootstrap je fakt, že nepotřebujeme znát druh výběrového rozdělení. Metodu je tedy možné aplikovat na každý náhodný výběr.

Pro porozumění metodě bootstrap si představme, že z náhodné veličiny můžeme provést více náhodných výběrů stejné délky. Pak z těchto výběrů získáme výběrové rozdělení a na jeho základě získáme parametry, které chceme zjistit. Tento postup je však nepraktický, jelikož opakované provádění náhodného výběru z náhodné veličiny může být nemožné, nebo velmi nákladné (časově i zdrojově; navíc bychom tímto postupem ztratili smysl výběrové statistiky). Metoda bootstrap umožňuje tyto nedostatky eliminovat. Považujme náhodný výběr za novou populaci. Z této populace provádíme náhodné výběry s opakováním. Tyto výběry nazveme **bootstrapové výběry**. Bootstrapových výběrů děláme většinou několik tisíc (doporučeno je minimálně tisíc). Z jednotlivých bootstrapových výběrů poté můžeme odhadnout parametr náhodné veličiny, který chceme znát (podrobněji se tímto budeme zabývat dále v kapitole).[6]

Poznámka 3.1 Bootstrapový výběr (náhodný výběr s opakováním) je kombinace s opakováním. Celkový počet unikátních výběrů délky n je $\binom{2n-1}{n}$.

Podívejme se nyní na vzorový příklad, jak provedeme bootstrapový výběr:

Příklad 1: Mějme náhodný výběr $X_v = \{1, 2, 3, 4, 5\}$, pro který chceme určit bootstrapové výběry. Prvních 5 bootstrapových výběrů určíme například jako:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	3	2	5	5	4
[2,]	2	1	2	1	2
[3,]	3	4	2	2	5
[4,]	3	1	1	1	3
[5,]	2	2	5	3	4

Výpis 1: Výstup generování bootstrapových výběrů

Vidíme, že v jednotlivých bootstrapových výběrech se mohou prvky náhodného výběru opakovat (např. v prvním řádku je dvakrát prvek 5), nebo se nemusí vyskytovat vůbec (např. v prvním řádku není prvek 1). Pokud budeme simulaci opakovat, dostaneme

jiné bootstrapové výběry (což je dáno prováděním náhodných výběrů s opakováním). Další postup, poté co získáme požadovaný počet bootstrapových výběrů, závisí na námi zvolené bootstrapové metodě. Blíže se podíváme na kvantilovou bootstrapovou metodu, která je dnes nejrozšířenější.

3.1 Kvantilový bootstrapový interval

Kvantilová bootstrapová metoda (angl. bootstrap percentile method) je dnes nejrozšířenější a nejpopulárnější díky své jednoduchosti a účinnosti. Nejčastěji sestavované intervalové odhady mají spolehlivost $1-\alpha$ rovnu 95% nebo 99%. V následnicích případech budeme používat pouze spolehlivost $1-\alpha = 95\%$. Mějme $1\ 000$ bootstrapových hodnot parametru $\hat{\Theta}$, značených $(\hat{\Theta}^*_1, \hat{\Theta}^*_2, ..., \hat{\Theta}^*_{1\ 000})$. Pro sestavení 95% intervalu spolehlivosti vezmeme prvek na pozici 25 jako dolní mez a prvek na pozici 975 jako horní mez. Interval spolehlivosti tedy v tomto případě bude mít tvar:

$[\Theta^*_{25}, \Theta^*_{975}]$.

Postup při určování intervalového odhadu parametru Θ pomocí jednoduchého kvantilového IO z náhodného výběru $X_1, X_2, ..., X_k$ lze shrnout následujícím algoritmem [4]:

1. Určíme B bootstrapových výběrů o rozsahu n. Vybíráme z pozorovaných hodnot $(x_1, x_2, ..., x_n)$ náhodného výběru $(X_1, X_2, ..., X_n)$.
2. Vypočítáme odhad $\hat{\Theta}^*$ parametru Θ pro každý bootstrapový výběr.
3. Všechny vypočítané odhady parametru seřadíme vzestupně podle velikosti $\hat{\Theta}^*_1, \hat{\Theta}^*_2, ..., \hat{\Theta}^*_n$.
4. $\frac{\alpha}{2}$ a $\frac{1-\alpha}{2}$ kvantily odhadneme co nejpřesněji hodnotami $\hat{\Theta}^*_{\frac{\alpha}{2}}$ a $\hat{\Theta}^*_{\frac{1-\alpha}{2}}$ tak, aby platilo:

$$\frac{|\{\hat{\Theta}^*; \hat{\Theta}^* \leq \hat{\Theta}^*_{\frac{\alpha}{2}}\}|}{B} \doteq \frac{\alpha}{2}, \qquad \frac{|\{\hat{\Theta}^*; \hat{\Theta}^* \leq \hat{\Theta}^*_{\frac{1-\alpha}{2}}\}|}{B} \doteq \frac{1-\alpha}{2}.$$

Jednoduchým kvantilovým intervalem spolehlivosti pro parametr Θ se spolehlivostí $1-\alpha$ nazveme interval: $\langle \hat{\Theta}^*_{\frac{\alpha}{2}}; \hat{\Theta}^*_{\frac{1-\alpha}{2}} \rangle$.

Poznámka 3.2 Velikost B v prvním kroku by se měla volit alespoň $1\ 000$.

$|\{\}|$ značí velikost množiny (počet jejich prvků).

Podívejme se nejprve na jednoduchý příklad sestavení intervalového odhadu kvantilovým bootstrapem. Jako náhodný výběr nám poslouží data z příkladu u bootstrapového výběru.

Příklad 2: Mějme náhodný výběr $X_v = \{1, 2, 3, 4, 5\}$ z neznámé náhodné veličiny X_n. Pro tento výběr chceme určit 95% intervalový odhad střední hodnoty. Zvolme počet opakování bootstrapu $B = 1000$. Nejprve tedy určíme $1\ 000$ náhodných výběrů s opakováním z X_v:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	3	1	4	2	2
[2,]	2	3	1	3	3
[3,]	5	2	4	2	5

```
          ...  ...  ...  ...  ...  ...
[999,]      5    1    1    2    4
[1000,]     4    2    2    5    1
```

Výpis 2: Ukázka z $1\ 000$ bootstrapových výběrů

Pro každý bootstrapový výběr následně určíme hodnotu jeho výběrového průměru:

```
[1]    2.4
[2]    2.4
[3]    3.6
...    ...
[999]  2.6
[1000] 2.8
```

Výpis 3: Ukázka z $1\ 000$ výběrových průměrů

Výběrové průměry následně seřadíme vzestupně podle velikosti a určíme v pořadí $25.$ a $975.$ hodnotu ($1\ 000 \cdot \frac{0,05}{2}$ a $1\ 000 \cdot (1 - \frac{0,05}{2})$) jako dolní a horní mez 95% intervalového odhadu střední hodnoty: $(1,6; 4,2)$.

Jako demonstraci použití kvantilového odhadu se podíváme na praktický příklad [7].

Příklad 3: Chceme určit hodnotu IQ populace ČR starší 18 let. Změřili jsme IQ u 20 náhodně vybraných lidí. Výsledky máme ve vektoru (61, 88, 89, 89, 90 92, 93, 94, 98, 98, 101, 102, 105, 108, 109, 113, 114, 115, 120, 138). Určit bodový odhad střední hodnoty (výběrový průměr) pro tato data není problém (výběrový průměr: $100,9$). Většinou nás však nezajímá bodový odhad, nýbrž odhad intervalový. Nejprve si představme, že hodnoty IQ mají normální rozdělení. V takovém případě je sestavení intervalového odhadu střední hodnoty možno provést pomocí statistického software či statistických tabulek. Předpokládáme, že směrodatnou odchylku neznáme. Interval spolehlivosti tedy sestavíme pomocí 2.6.2. V našem případě je 95% interval spolehlivosti odhadu střední hodnoty: $(93,4; 108,3)$.

Jenže nevíme jistě, zda námi naměřené hodnoty IQ skutečně pocházejí z normálního rozdělení. V tomto případě tedy bude lepší se uchýlit k metodě bootstrap.

Obdobně jako v předchozím příkladu, budeme provádět kombinace s opakováním, tentokrát délky 20. Pro každou následně určíme výběrový průměr. Tentokrát provedeme $100\ 000$ bootstrapových výběrů.

Bootstrapový výběr $\#1$: 61, 88, 88, 89, 89, 90, 92, 93, 98, 102, 105, 105, 105, 109, 109, 109, 109, 114, 114, 120. Výběrový průměr $= 99,5$.

Bootstrapový výběr $\#2$: 61, 88, 89, 89, 90, 92, 92, 98, 98, 98, 102, 105, 105, 108, 108, 113, 113, 113, 114, 138. Výběrový průměr $= 100,7$.

...

Bootstrapový výběr $\#99\ 999$: 61, 61, 88, 89, 92, 93, 93, 94, 98, 98, 98, 101, 102, 105, 109, 114, 115, 120, 120, 138. Výběrový průměr $= 99,5$.

Bootstrapový výběr $\#100\ 000$: 61, 61, 61, 88, 89, 89, 90, 93, 93, 94, 102, 105, 108, 109, 109, 114, 115, 115, 120, 138. Výběrový průměr $= 97,7$.

Následně vektor výběrových průměrů seřadíme vzestupně a jako dolní mez určíme prvek na pozici $100\ 000 \cdot \frac{0,05}{2} = 2\ 500$, jako horní mez určíme prvek na pozici $100\ 000 \cdot (1 - \frac{0,05}{2}) = 97\ 500$. Konkrétně je intervalový odhad střední hodnoty: $(94,0; 107,6)$.

Podívejme se blíže na intervalové odhady střední hodnoty. Nejprve jsme předpokládali, že rozdělení IQ je normální a příslušný interval jsme spočítali. Poté jsme intervalový odhad sestavili pomocí kvantilového bootstrapu. V následující tabulce je jejich srovnání:

rozsah výběru $n = 20$	dolní mez	horní mez
klasický IO	$93,4$	$108,3$
bootstrapový IO $B = 1\ 000$	$94,6$	$107,6$
bootstrapový IO $B = 10\ 000$	$94,0$	$107,5$
bootstrapový IO $B = 100\ 000$	$94,0$	$107,6$

Obrázek 1: Porovnání šířky intervalových odhadů střední hodnoty IQ občanů ČR

V tomto případě je bootstrapový interval užší. Jak se však ukáže, ne vždy tomu tak je. Další zajímavé srovnání je vliv počtu opakování bootstrapových simulací (B) na velikost intervalu. Vidíme, že již pro $1\ 000$ simulací dostaneme velmi podobné výsledky jako pro $100\ 000$ opakování. V další části práce tedy budeme provádět $1\ 000$ bootstrapových simulací.

4 Porovnávání efektivity

V této kapitole se bude porovnávat efektivita bootstrapového intervalového odhadu oproti klasickému výpočtu intervalového odhadu (IO). Pro porovnávání použijeme pravděpodobnost pokrytí intervalového odhadu. V následujících simulacích budeme používat náhodně generovaná data z náhodných veličin. Pro jejich vygenerování a pro určení pravděpodobnosti pokrytí intervalového odhadu potřebujeme znát parametry náhodné veličiny. Z tohoto důvodu budeme sestavovat intervalové odhady parametrů, jejichž skutečnou hodnotu známe.

Předpokládáme, že pravděpodobnost pokrytí bude blízká zvolené spolehlivosti $1 - \alpha = 0,95$. Jak se však ukáže, není tomu vždy tak.

4.1 Testování

Jedním z výstupů této práce je porovnání efektivity bootstrapového intervalového odhadu oproti klasickým metodám stanovení intervalového odhadu. V této kapitole budeme za klasický způsob stanovení intervalového odhadu považovat postup jako v 2.6.2. K určení efektivity se bude stanovovat pravděpodobnost pokrytí (CP, z angl. coverage probability) intervalového odhadu. Před definováním samotné pravděpodobnosti pokrytí budeme potřebovat pomocný identifikátor I. Mějme parametr Θ, který odhadneme intervalovým odhadem (IO). Pak:

Definice 4.1 *Identifikátor I pro neznámý parametr Θ nabývá hodnot:*

$$I = \begin{cases} 1, & \Theta \in IO \\ 0, & \Theta \notin IO \end{cases}$$

Identifikátor tedy nabývá hodnoty 1, pokud odhadnutý parametr skutečně leží v intervalovém odhadu.

Mějme M intervalových odhadů, pro které jsme určili hodnotu identifikátoru I. Pak pravděpodobnost pokrytí určíme jako:

Definice 4.2 *Pravděpodobnost pokrytí se definuje jako [9]:*

$CP(n, \mu, \sigma, 1 - \alpha) = P(\Theta \in IO) = \sum_{i=1}^{M} I_i / M$

M v definici značí počet opakování testu. Mějme náhodnou veličinu se střední hodnotou μ, směrodatnou odchylkou σ. Zvolme počet opakování testu $M = 1\ 000$, spolehlivost odhadu $1-\alpha = 0,95$. Dále víme, že náhodná veličina, z které provádíme náhodné výběry, má normální rozdělení. Chceme určit pravděpodobnost pokrytí intervalového odhadu střední hodnoty. Pak provedeme $1\ 000$ náhodných výběrů o rozsahu n, pro které sestavíme intervalové odhady střední hodnoty. Porovnáme, jestli skutečná hodnota střední hodnoty v intervalu leží a podle toho určíme hodnotu identifikátoru I. Očekáváme, že pro hladinu spolehlivosti $1-\alpha = 0,95$ bude přibližně 950 intervalů obsahovat skutečnou hodnotu μ. Následně sečteme hodnoty identifikátoru I pro každý intervalový odhad a podělíme celkovým počtem opakování testu M. Výsledek je číslo z intervalu $(0; 1)$, které udává pravděpodobnost pokrytí intervalového odhadu.

4.2 Normální rozdělení

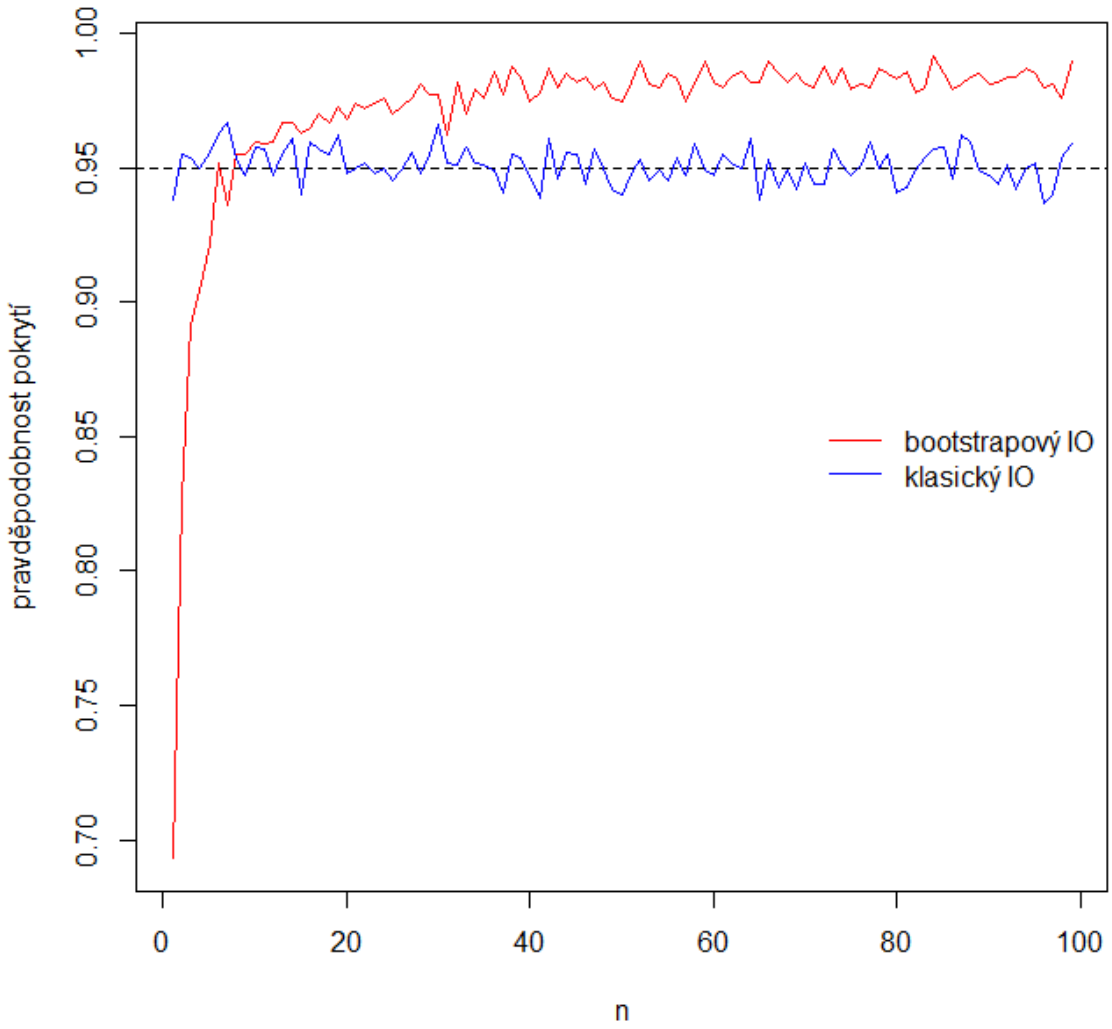
Vraťme se k příkladu **3** z kapitoly Bootstrap. Zde jsme měli k dispozici 20 hodnot IQ u náhodně vybraných lidí z populace ČR. Stanovit pro tento příklad pravděpodobnost pokrytí intervalového odhadu by vyžadovalo $1\ 000$ různých náhodných výběrů hodnot IQ z populace ČR (pokud bychom počet opakování testu M volili $1\ 000$). Získat tolik náhodných výběrů (např. změřením IQ u $20\ 000$ náhodně vybraných občanů ČR) by bylo velmi nákladné. Předpokládejme proto, že IQ občanů ČR má normální rozdělení se střední hodnotou $\mu = 100$ a směrodatnou odchylkou $\sigma = 10$. S pomocí těchto parametrů jsme schopni náhodné výběry počítačově simulovat. Jejich délku budeme značit n.

Při sestavení klasického intervalového odhadu střední hodnoty budeme postupovat jako v 2.6.2. Nejprve vygenerujeme náhodný výběr délky n. Následně sestavíme interval spolehlivosti. Předpokládáme, že hodnotu směrodatné odchylky neznáme (jelikož v praxi je směrodatná odchylka často neznámá). Porovnáme, jestli skutečná střední hodnota IQ ($\mu = 100$) leží v sestaveném intervalu spolehlivosti. Uložíme si hodnotu identifikátoru I a vygenerujeme další náhodný výběr. Celý postup opakujeme tisíckrát. Poté z hodnot identifikátoru I sestavíme pravděpodobnost pokrytí intervalového odhadu střední hodnoty.

Při sestavování kvantilového bootstrapového intervalového odhadu se postupuje obdobně. Vygenerujeme náhodný výběr délky n. Pomocí kvantilového bootstrapu sestavíme interval spolehlivosti a porovnáme, zda skutečná hodnota střední hodnoty v něm leží. Určíme hodnotu identifikátoru I a uložíme. Provedeme další náhodný výběr. Celý postup opakujeme tisíckrát a poté určíme pravděpodobnost pokrytí intervalového odhadu střední hodnoty.

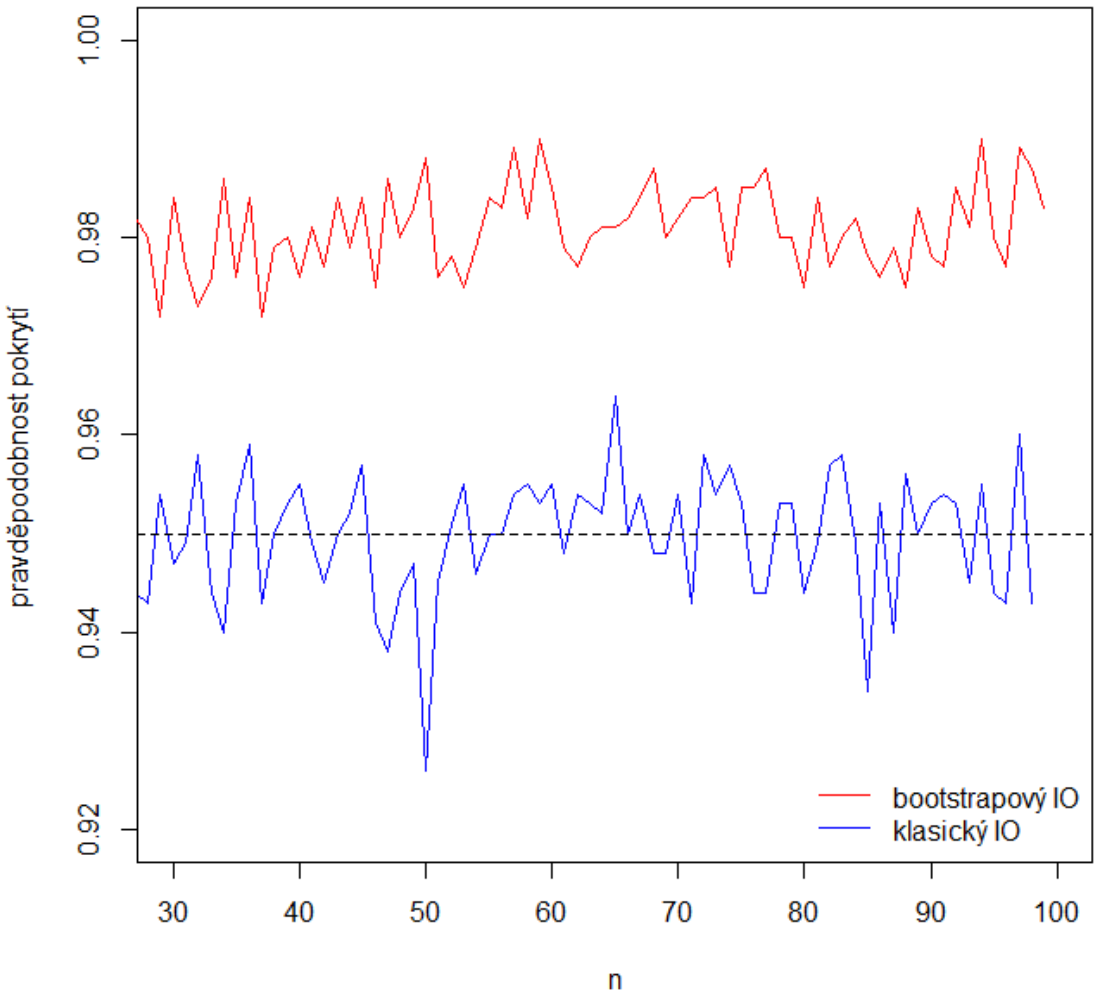
Nejprve se podíváme na porovnání pravděpodobnosti pokrytí intervalového odhadu pro normální rozdělení $N(100, 10)$.

Obrázek 2: Pravděpodobnost pokrytí pro data pocházející z normálního rozdělení $N(100, 10)$ v závislosti na změně délky náhodných výběrů n z intervalu $\langle 2, 100 \rangle$.



Z grafu je patrné, že pro malou délku náhodných výběrů $n (n < 10)$ metoda bootstrap dosahuje malé pravděpodobnosti pokrytí. Toto chování způsobuje právě malá délka náhodného výběru, ze kterého bootstrapové výběry provádíme. Oproti tomu klasický způsob sestavení intervalu spolehlivosti nabývá i pro malé délky náhodných výběrů pravděpodobnosti pokrytí přibližně $0,95$. Toto je v souladu s naším očekáváním, jelikož jsme hodnotu spolehlivosti $1 - \alpha$ zvolili $0,95$. Podívejme se dále na detailnější část grafu pro $n \geq 30$:

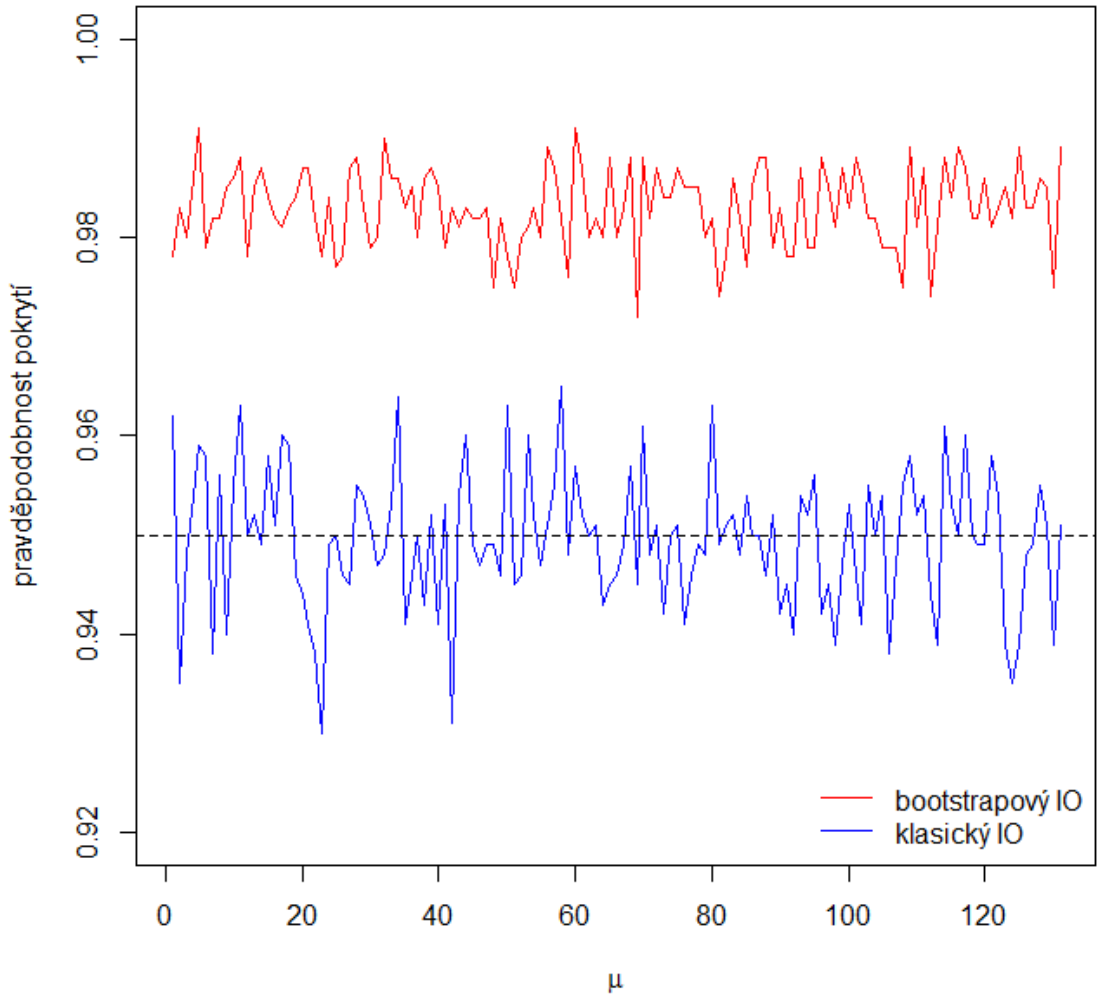
Obrázek 3: Pravděpodobnost pokrytí pro data pocházející z normálního rozdělení $N(100, 10)$ v závislosti na změně délky náhodných výběrů n z intervalu $\langle 30, 100 \rangle$.



Vidíme, že při použití kvantilové bootstrapové metody dosahujeme vyšší pravděpodobnosti pokrytí intervalů spolehlivosti než je požadovaná hodnota spolehlivosti $0,95$. Takovéto intervaly se nazývají konzervativní.

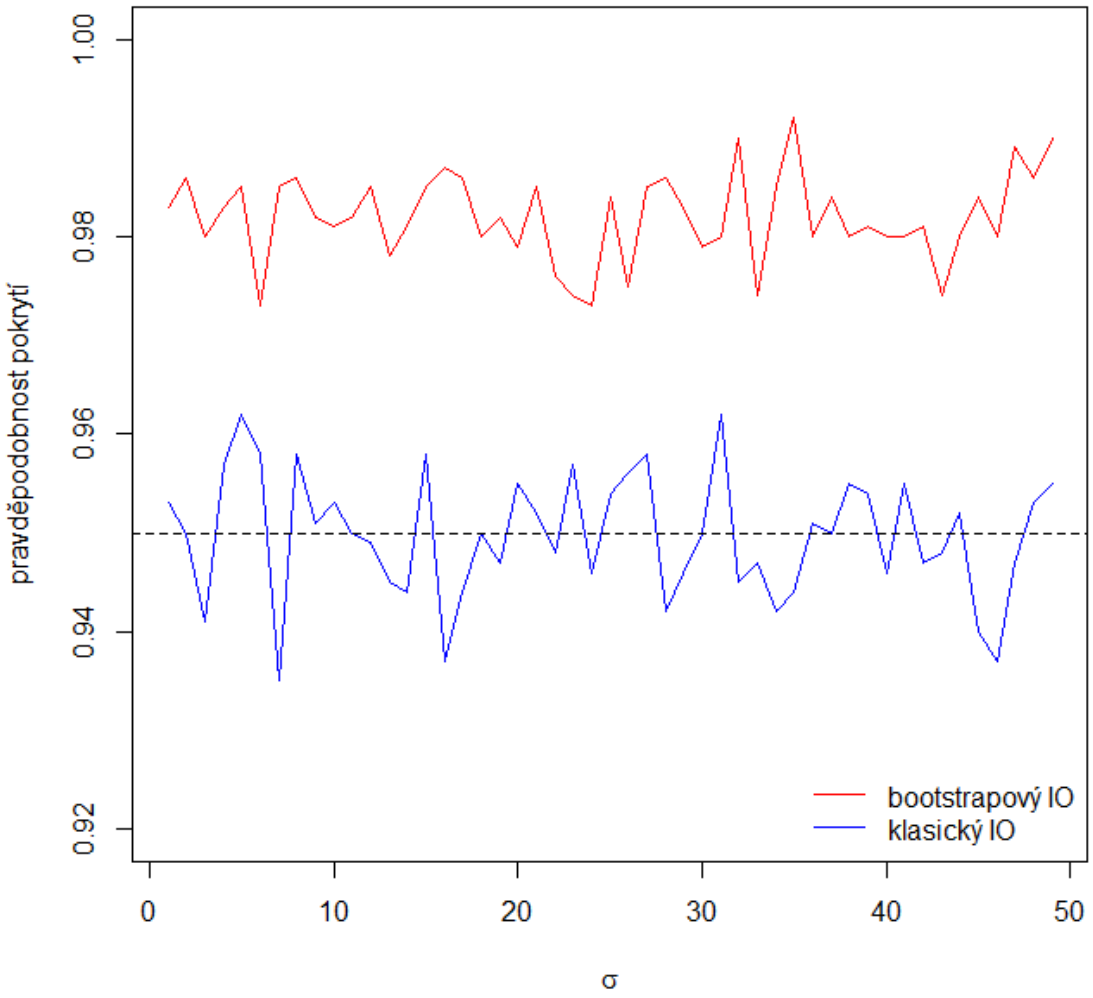
Dále se podíváme, co se spolehlivostí bude dít při změně střední hodnoty μ, resp. směrodatné odchylky σ (počet prvků bude konstantní $n = 100$). Nejprve se podíváme na graf pro změnu střední hodnoty μ (směrodatná odchylka je konstantní $\sigma = 10$):

Obrázek 4: Pravděpodobnost pokrytí pro data pocházející z normálního rozdělení $N(100, 10)$ v závislosti na změně střední hodnoty μ z intervalu $\langle 1, 130 \rangle$.



Z grafu je patrné, že při změně střední hodnoty μ se bootstrapový interval spolehlivosti i klasický interval spolehlivosti chová obdobně, jako v předchozím příkladě (bootstrapový intervalový odhad je konzervativní; klasický intervalový odhad má pravděpodobnost pokrytí rovnu přibližně $0,95$). Velikost pravděpodobnosti pokrytí se pro střední hodnotu μ z intervalu $< 1, 130 >$ nemění. Změna střední hodnoty tedy neovlivní velikost pravděpodobnosti pokrytí intervalového odhadu. Podívejme se nyní na změnu směrodatné odchylky σ (střední hodnota je konstantní $\mu = 100$):

Obrázek 5: Pravděpodobnost pokrytí pro data pocházející z normálního rozdělení $N(100, 10)$ v závislosti na změně směrodatné odchylky σ z intervalu $\langle 1, 50 \rangle$.

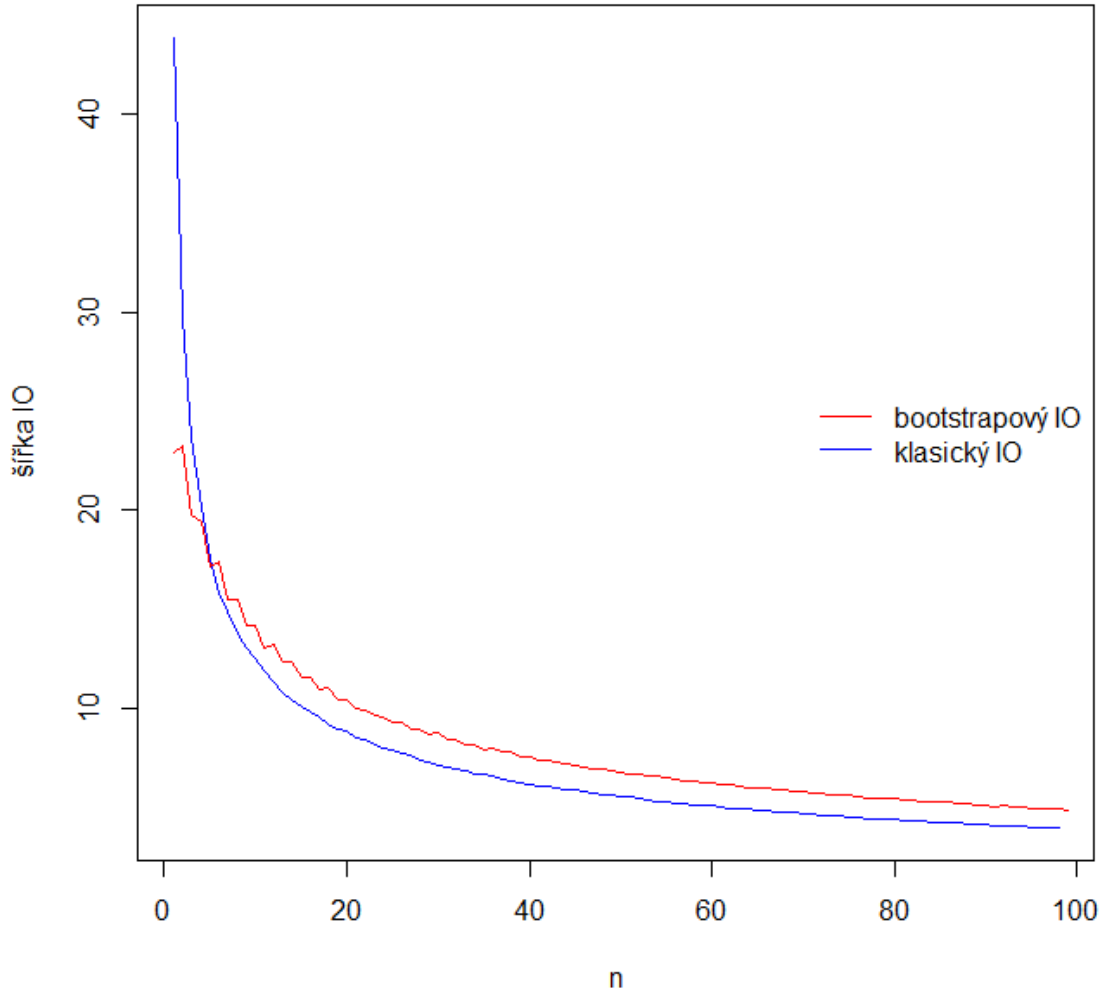


Obdobně jako v předchozím případě (při změně střední hodnoty μ), při změně směrodatné odchylky σ je bootstrapový intervalový odhad konzervativní a klasický intervalový odhad dosahuje pravděpodobnosti pokrytí přibližně $0,95$. Ani změna směrodatné odchylky tedy nemá vliv na pravděpodobnost pokrytí intervalového odhadu.

Z předchozích simulací je patrné, že pro normální rozdělení stanovuje kvantilová bootstrapová metoda konzervativní intervalové odhady (tedy odhady, jejichž spolehlivost je vyšší než požadovaná). Avšak takto sestavený interval by měl být pro spolehlivost $1 - \alpha = 0,95$ širší v porovnání s klasickým intervalovým odhadem. Toto nyní ověříme. V obou metodách se pro různé délky náhodných výběrů n z intervalu $< 1, 100 >$ sestaví $1\ 000$ intervalových odhadů. Z těchto intervalů určíme průměrnou dolní a horní mez.

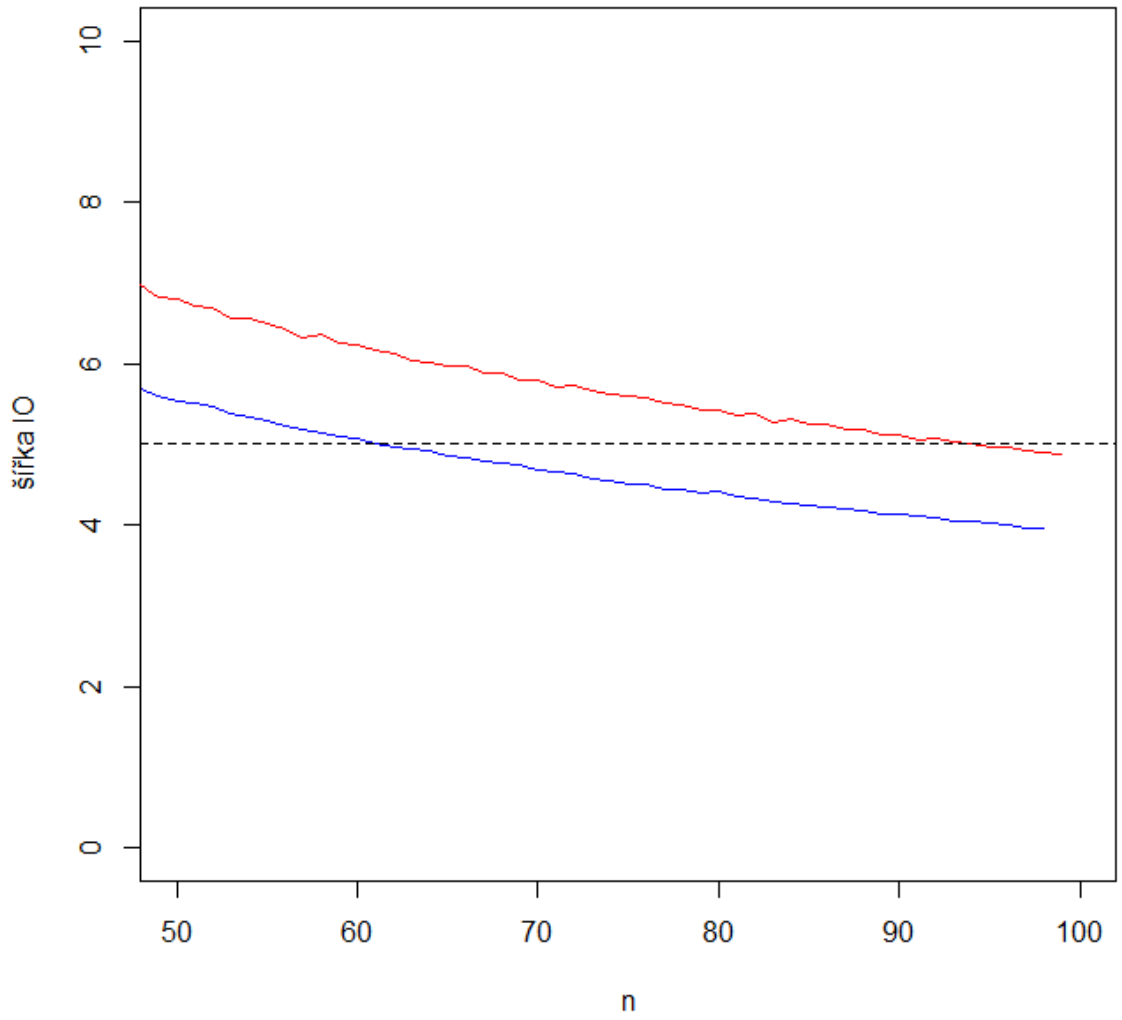
Dolní mez odečteme od horní, čímž získáme šířku intervalu. Podívejme se na výsledný graf:

Obrázek 6: Průměrná šířka intervalového odhadu pro data z normálního rozdělení $N(100, 10)$ v závislosti na změně délky náhodných výběrů n z intervalu $\langle 2, 100 \rangle$.



Z grafu je patrné, že pro velmi malou délku náhodných výběrů n sestavuje klasická metoda velmi široké intervalové odhady. Ale již pro $n > 10$ je šířka klasických intervalových odhadů menší při porovnání s kvantilovou bootstrapovou metodou. Podívejme se detailněji na graf pro délku náhodných výběru $n \geq 50$:

Obrázek 7: Průměrná šířka intervalového odhadu pro data z normálního rozdělení $N(100, 10)$ v závislosti na změně délky náhodných výběrů n z intervalu $\langle 50, 100 \rangle$.

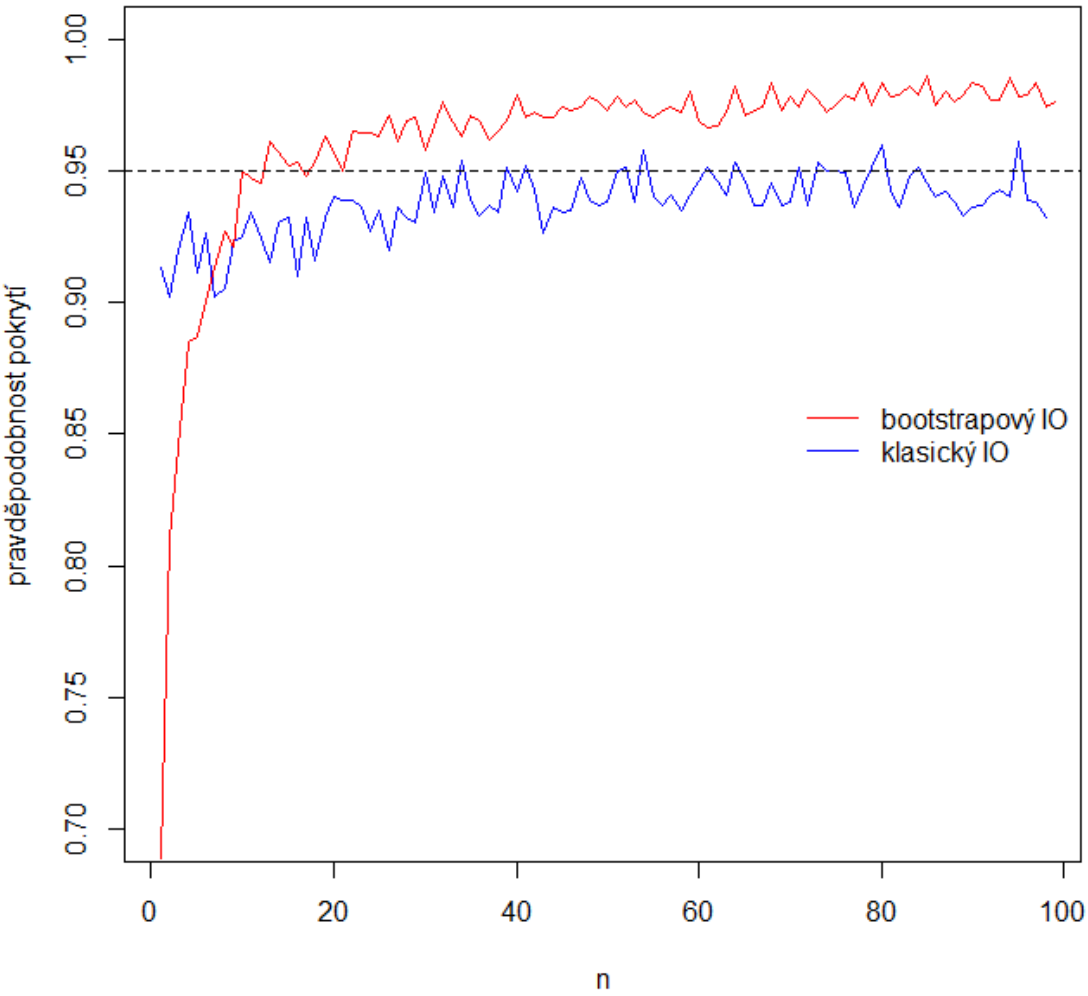


Z grafů je patrná výhoda klasického sestavování intervalového odhadu. Intervalový odhad sestavený klasicky je užší než intervalový odhad sestavený pomocí kvantilové bootstrapové metody. Intervalový odhad sestavený klasicky tedy dosahuje hodnoty pravděpodobnosti pokrytí intervalového odhadu přibližně rovné požadované spolehlivosti. A stanoví užší intervalový odhad (v porovnání s kvantilovým bootstrapem pro stejnou délku náhodného výběru n). Interval spolehlivosti stanovený klasicky je v tomto případě asi o 17% užší, než kvantilový bootstrapový interval spolehlivosti pro stejnou délku náhodného výběru n, kde $n \geq 30$. Jenže použití klasické metody podmiňuje splnění podmínek (viz předpoklady pro 2.6.2). Jak si ukážeme dále, nesprávné použití klasického intervalového odhadu povede ke špatným výsledkům.

4.3 Lognormální rozdělení

Při porovnávání efektivity budeme postupovat obdobně jako u normálního rozdělení. Důležité je však zdůraznit, že použití klasického intervalového odhad je nekorektní a tento intervalový odhad je použit pouze pro možnost porovnání. Nejprve se podíváme na chování intervalových odhadů při změně délky náhodného výběru pro lognormální rozdělení s parametry $LN(100, 10)$:

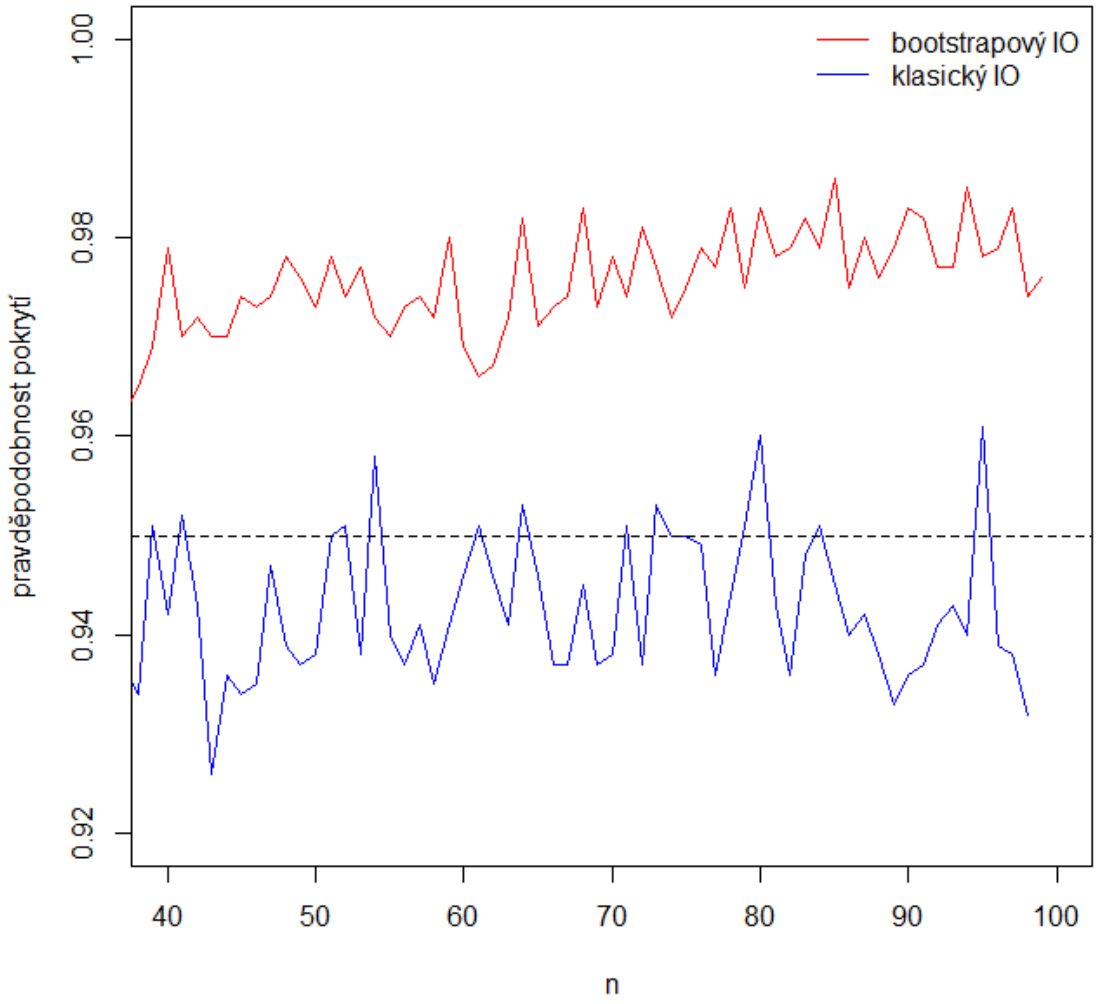
Obrázek 8: Pravděpodobnost pokrytí pro data z lognormálního rozdělení $LN(100, 10)$ v závislosti na délce náhodného výběru n z intervalu $\langle 2, 100 \rangle$.



Z grafu je patrné, že klasický výpočet intervalového odhadu ve většině případů pravděpodobnosti pokrytí menší než $0,95$. Kvantilový bootstrapový intervalový odhad pro

velmi malá n dosahuje nízkého koeficientu spolehlivosti, obdobně jako u normálního rozdělení. Podívejme se dále na část grafu pro délku náhodného výběru $n > 40$:

Obrázek 9: Pravděpodobnost pokrytí pro data z lognormálního rozdělení $LN(100, 10)$ v závislosti na délce náhodného výběru n z intervalu $\langle 40, 100 \rangle$.



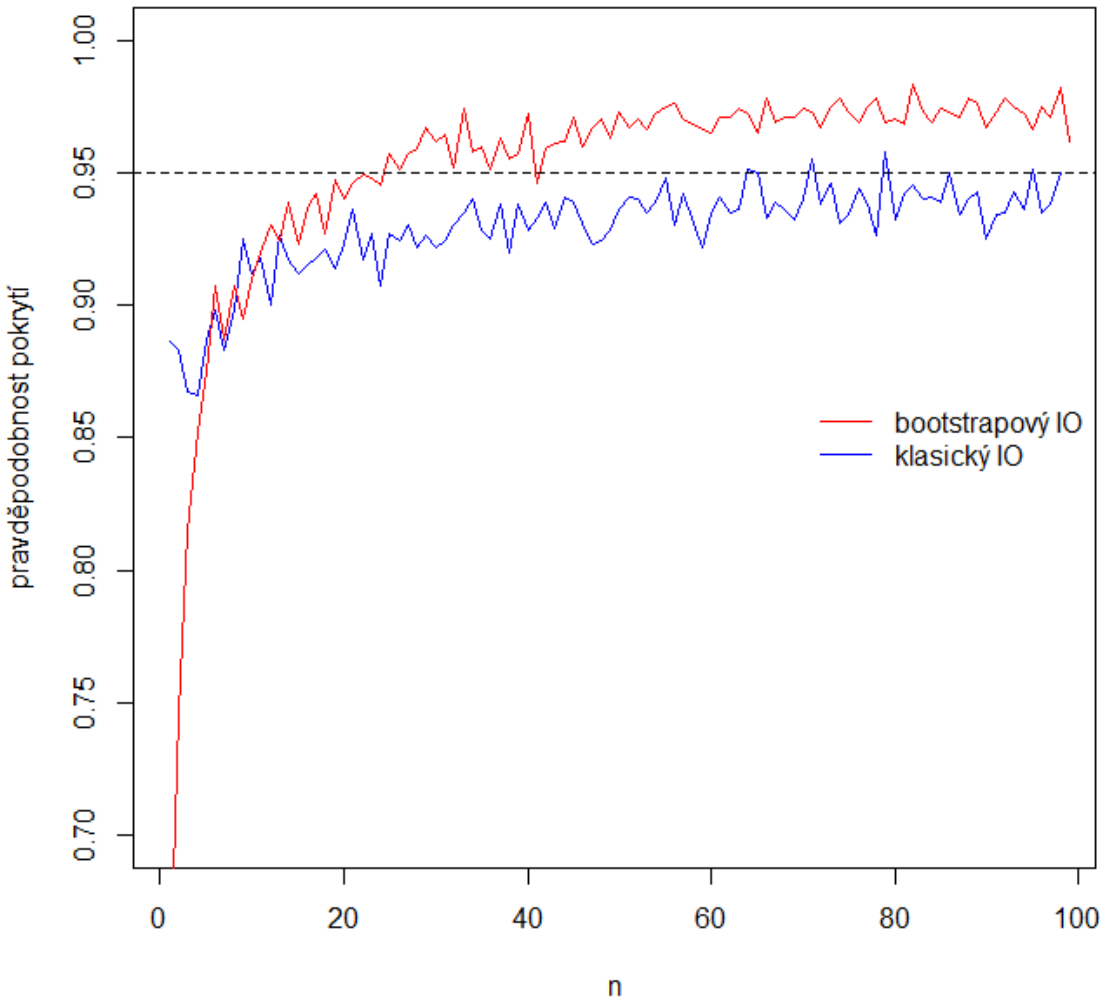
Nekorektně použitá klasická metoda má pravděpodobnost pokrytí menší než stanovenou spolehlivost ($1-\alpha = 0,95$). Kvantilový bootstrapový intervalový odhad stanovuje konzervativní odhady. Jeho průběh je podobný, jako u normálního rozdělení.

Podívejme se dále na chování intervalových odhadů pro exponenciální rozdělení.

4.4 Exponenciální rozdělení

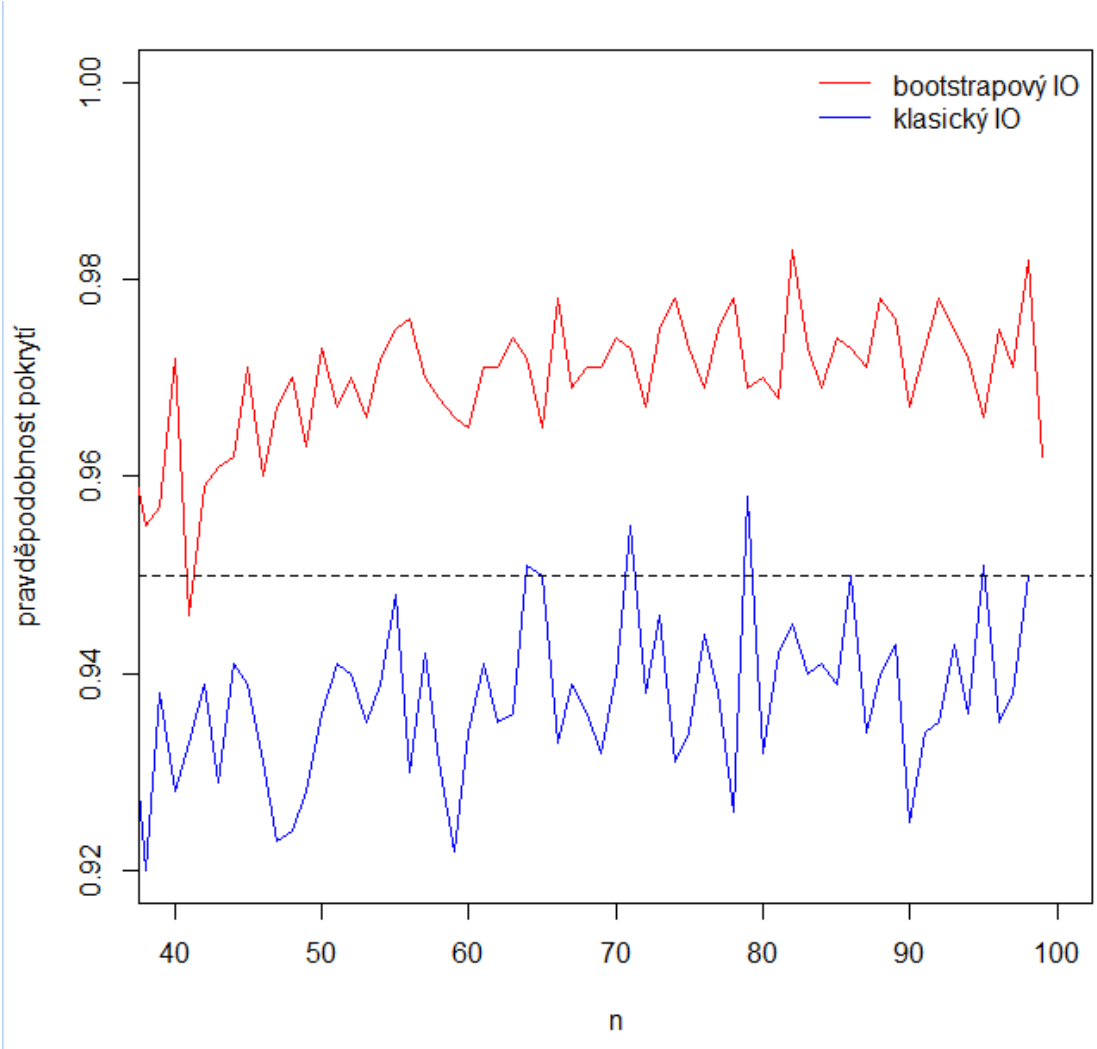
Při porovnávání efektivity pro exponenciální rozdělení budeme postupovat obdobně jako u normálního rozdělení. Je však důležité zdůraznit, že klasický způsob určení intervalového odhadu je nekorektní, obdobně jako u lognormálního rozdělení. Podívejme se na graf změny délky náhodných výběrů n pro exponenciální rozdělení $Exp(1)$:

Obrázek 10: Pravděpodobnost pokrytí pro data z exponenciálního rozdělení $Exp(1)$ v závislosti na délce náhodného výběru n z intervalu $\langle 2, 100 \rangle$.



Klasická metoda nedosahuje pravděpodobnosti pokrytí intervalového odhadu $0,95$. Kvantilový bootstrapový intervalový odhad pro malé délky náhodných výběrů n taktéž. Podívejme se detailněji na graf pro délku náhodných výběrů $n > 40$:

Obrázek 11: Pravděpodobnost pokrytí pro data z exponenciálního rozdělení $Exp(1)$ v závislosti na délce náhodného výběru n z intervalu $\langle 40, 100 \rangle$.



Kvantilový bootstrapový intervalový odhad má průběh podobný jako u normálního a lognormálního rozdělení. Klasický intervalový odhad nedosahuje pravděpodobnosti pokrytí $0,95$, obdobně jako u lognormálního rozdělení. Toto je způsobeno jeho nekorektním použitím.

V následující kapitole si shrneme, co jsme o kvantilovém bootstrapovém intervalovém odhadu zjistili.

4.5 Vyhodnocení

V předchozích simulacích jsme zjistili, že kvantilový bootstrap má podobné chování pro všechny 3 druhy rozdělení. Skutečně tedy nezáleží na tom, z jakého rozdělení data po-

cházejí. Dále jsme ukázali, že nesprávné použití klasické metody vede k intervalům spolehlivosti, jejichž pravděpodobnost pokrytí je nižší, než požadovaná spolehlivost.

Naše poznatky můžeme shrnout následovně:

- Kvantilový bootstrapový intervalový odhad stanovuje takzvané konzervativní intervalové odhady. Konzervativní intervalový odhad má pravděpodobnost pokrytí vyšší než je požadovaný koeficient spolehlivosti.
- Kvantilový bootstrapový intervalový odhad stanovuje širší intervalový odhad (oproti klasickému intervalovému odhadu). Širší intervalový odhad má větší rozdíl mezi horní a dolní mezí. Tato vlastnost přímo souvisí s konzervativností bootstrapu - vyšší pravděpodobnost pokrytí vyžaduje větší počet prvků, které v něm musí ležet.
- Kvantilový bootstrapový intervalový odhad má podobný průběh pro normální, lognormální i exponenciální rozdělení. Toto souvisí s velkou výhodou kvantilového bootstrapu - jeho fungování je nezávislé na druhu rozdělení.
- Klasické stanovování intervalových odhadů je nevhodné pro data, pro která nejsme schopni zaručit předpoklady k použití klasické metody.

Poznámka 4.1 Souvislost mezi konzervativním intervalovým odhadem a jeho šířkou lze pozorovat v grafech závislosti normálního rozdělení na délce náhodných výběrů a závislosti šířky intervalů na délce náhodných výběrů. V nich jde vidět, že pro velmi malá n je pravděpodobnost pokrytí bootstrapového intervalového odhadu menší a zároveň šířka intervalového užší (při porovnání s klasickým intervalovým odhadem). Poté dojde ve stejném n k protnutí grafů a záměně těchto vlastností (tedy bootstrapový intervalový odhad má větší pravděpodobnost pokrytí a širší intervalový odhad). Toto není náhoda, jelikož pravděpodobnost pokrytí intervalového odhadu a šířka intervalového odhadu spolu souvisí (čím větší pravděpodobnost pokrytí, širší intervalový odhad je).

5 Závěr

Cílem bakalářské práce bylo seznámit se s metodou bootstrap a prozkoumat její možnosti při konstrukci intervalů spolehlivosti. Nejprve jsme si připomněli některé statistické pojmy. V následující kapitole jsme popsali použití metody bootstrap při konstrukci intervalů spolehlivosti. Způsobů konstrukce je několik, my se však zaměřili na nejznámější - kvantilovou. V poslední kapitole jsme pak porovnávali, jaké vlastnosti mají intervaly spolehlivosti sestavené kvantilovým bootstrapem oproti intervalům spolehlivosti sestavené klasicky. Zde se ukázal rozdíl, který platil pro všechny mnou simulované situace: interval spolehlivosti sestavený kvantilovou bootstrapovou metodou je konzervativní, ale širší. Dále jsme si ukázali, že kvantilový bootstrap je použitelný nezávisle na druhu rozdělení, ze kterého náhodný výběr pochází. Navíc jsme si ukázali úskalí nerozvážného použití klasického sestavení intervalového odhadu.

Avšak metoda bootstrap se zdaleka neomezuje pouze na problematiku konstrukce intervalů. Jedná se o velmi univerzální metodu se širokou škálou použití. její uplatnění je mimo jiné také v bayesovské statistice, při testování hypotéz atd. Možností k dalšímu studování metody je tedy spousta. V přímé návaznosti na tuto práci je možnost prozkoumat jiné bootstrapové metody pro konstrukci intervalů spolehlivosti,jako je t-interval pomocí bootstrapu či BCa bootstrap (což je metoda vycházející z kvantilového bootstrapu).

6 Reference

[1] ANDĚL, Jiří. *Základy matematické statistiky.* Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-7378-162-0.

[2] MANN, Prem S. *Introductory statistics.* 8th ed., international student version. Singapore: Wiley, c2013. ISBN 978-1-118-31870-6.

[3] PRÁŠKOVÁ, Zuzana. Metoda bootstrap. In *Robust 2004. Sborník prací 13. letní školy JČMF ROBUST 2004 uspořádané Jednotou českých matematiků a fyziků za podpory KPMS MFF UK a České statistické společnosti ve dnech 7. – 11. června 2004 v Třešti.* Brno, 2004. s. 299-314. ISBN 80-7015-972-3.

[4] PAVLÍČKOVÁ, Lucie. *Metoda bootstrap a její aplikace.* Brno: Vysoké učení technické. Fakulta strojního inženýrství. Ústav matematiky, 2009. 63 s. Vedoucí diplomové práce doc. RNDr. Zdeněk Karpíšek, CSc.

[5] LITSCHMANNOVÁ, Martina. *Úvod do statistiky* [online]. Ostrava, 2011. Dostupné z: <http://mi21.vsb.cz/modul/uvod-do-statistiky>

[6] SINGH, Kesar, XIE, Minge. *Bootstrap: a statistical method* [online]. Ruthers university. Dostupné z: <http://stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf>

[7] PEZZULLO, John. *The Bootstrap Method for Standard Errors and Confidence Intervals* [online]. Dostupné z: <http://www.dummies.com/how-to/content/the-bootstrap-method-for-standard-errors-and-confi.html>

[8] BÍLKOVÁ, Diana, Petr BUDINSKÝ a Václav VOHÁNKA. *Pravděpodobnost a statistika.* Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2009. ISBN 978-80-7380-224-0.

[9] LITSCHMANNOVÁ, Martina. *Využití moderních statistických metod pro analýzu nežádoucích účinků spojených s radioterapií karcinomu prostaty.* Ostrava: Vysoké škola báňská. Fakulta elektrotechniky a informatiky. Katedra aplikované matematiky, 2011. 120 s. Vedoucí disertační práce Prof. Ing. Radim Briš, CSc.

A Zdrojové kódy

```
N=seq(1, 100, 1) ## Delka nahodneho vyberu n
nr = 1000 ## pocet opakovani simulace
nb = 1000 ## pocet bootstrapovych vyberu
CP = NULL ## vektor pravdepodobnosti pokryti
LI=NULL ## vektor dolnich mezi
PI=NULL ## vektor hornich mezi
for (j in 1:length(N))
{
n = N[j] ## ziskame delku nahodneho vyberu
nc = 0 ## ukladame si pocet IO, ve kterych skutecna hodnota lezi
lpom=0 ## ukladame si dolni mez
ppom=0 ## ukladame si horni mez

for (k in 1:nr)
{
X = rnorm(n,100,10) ##simulace nahodneho vyberu
mu = mean(X) ##vyberovy prumer
s = sd(X) ##vyberova smerodatna odchylka
pomoc = ceiling(n*runif(n*nb)) ##pripravim si nahodne vybery, ve vektoru mam cela cisla
B = X[pomoc] ## dosadim hodnoty misto cisel pozic ve vyberu (cislo ve vektru udava, ktery prvek
    z nahodneho vyberu vezmu)
B = array(B, c(nboot,n)) ##prevedu do jednorozmerneho pole
M = apply(B, 1, mean) ##ziskam stredni hodnoty pro kazdy bootstrapovy vyber
M = sort(M) ##setridim vzestupne stredni hodnoty
C = c(M[25], M[975]) ##ziskam meze
lpom=lpom+C[1]
ppom=ppom+C[2] ##ulozime dolni a horni mez
if ( (C[1] < 100) & (C[2] > 100) ) { nc = nc+1 } ##zkontrolujeme, zdali stredni hodnota lezi v
    intervalu
}
CP[j] = nc/nrep
LI[j]=lpom/nrep
PI[j]=ppom/nrep ##ulozime si hodnoty pravdepodobnosti pokryti a horni a dolni meze pro dalsi praci
    s nimi.
}
```

Výpis 4: Kvantilový bootstrap pro sestaveni IO střední hodnoty a pravděpodobnosti pokrytí IO